Easy Data

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam
Mathematical Institute – Leiden University

Joint work with
W. Koolen, T. Van Erven, N. Mehta, T. Sterkenburg
Today: Three Things To Tell You

1. **Nifty Reformulation** of Conditions for Fast Rates in Statistical Learning
   - Tsybakov, Bernstein, Exp-Concavity, ...

2. Do this via new concept: **ESI**

3. Precise Analogue of Bernstein Condition for Fast Rates in Individual Sequence Setting
   - ...and algorithm that achieves these rates!
Today: Three Things To Tell You

1. **Nifty Reformulation** of Conditions for **Fast Rates in Statistical Learning**

2. Do this via new concept: **ESI**

3. Precise Analogue of Bernstein Condition for **Fast Rates in Individual Sequence Setting**
   – ...and algorithm that achieves these rates!
Van Erven, G. Mehta, Reid, Williamson

Fast Rates in Statistical and Online Learning.

JMLR Special Issue in Memory of A. Chervonenkis, Oct. 2015

VC: Vapnik-Chervonenkis (1974!) optimistic (realizability) condition

TM: Tsybakov (2004) margin condition (special case: Massart Condition)


• Does not require 0/1 or absolute loss
• Does not require Bayes act to be in model
Decision Problem

• A decision problem (DP) is defined as a tuple \((P, \ell, \mathcal{F})\) where
  
  • \(P\) is the distribution of random quantity \(Z\) taking values in \(\mathcal{Z}\),
  
  • the **model** \(\mathcal{F}\) is a set of predictors \(f\), and for each \(f \in \mathcal{F}\), \(\ell_f : \mathcal{Z} \rightarrow \mathbb{R}\) indicates loss \(f\) makes on \(Z\)
  
  • **Example:** squared error loss

\[
Z = (X, Y) \\
f : \mathcal{X} \rightarrow \mathcal{Y} = \mathbb{R} \\
\ell_f(X, Y) = (Y - f(X))^2
\]
Decision Problem

- A decision problem (DP) is defined as a tuple \((P, \ell, \mathcal{F})\) where
  - \(P\) is the distribution of random quantity \(Z\) taking
    values in \(\mathcal{Z}\),
  - the **model** \(\mathcal{F}\) is a set of predictors \(f\), and for each
    \(f \in \mathcal{F}, \ell_f : \mathcal{Z} \to \mathbb{R}\) indicates loss \(f\) makes on \(Z\)
  - We assume throughout that the model contains a **risk minimizer** \(f^*\), achieving
    \[
    \mathbb{E}[\ell_{f^*}] = \inf_{f \in \mathcal{F}} \mathbb{E}[\ell_f]
    \]
  - \(\mathbb{E}[\ell_f]\) abbreviates \(\mathbb{E}_{Z \sim P}[\ell_f(Z)]\)
Bernstein Condition

• Fix a DP \((P, \ell, \mathcal{F})\) with (for now) \textbf{bounded} loss

• DP satisfies the \((C, \alpha)\)-Bernstein condition if there exists \(C > 0, \alpha \in [0,1]\), such that for all \(f \in \mathcal{F}\)

\[
\mathbb{E}[v_{f,f^*}] \leq C \cdot (\mathbb{E}[r_{f,f^*}])^\alpha
\]

where we set \(r_{f,f^*} = \ell_f - \ell_{f^*}\) and \(v_{f,f^*} = (r_{f,f^*})^2\)

• \(r_{f,f^*}\) is ‘regret of \(f\) relative to \(f^*\)’.
Bernstein Condition

• Fix a DP $\langle P, \ell, \mathcal{F} \rangle$ with (for now) bounded loss
• DP satisfies the $(C, \alpha)$-Bernstein condition if there exists $C > 0, \alpha \in [0,1]$, such that for all $f \in \mathcal{F}$

$$\mathbb{E}[v_{f,f^*}] \leq C \cdot (\mathbb{E}[r_{f,f^*}])^\alpha$$

where we set $r_{f,f^*} = \ell f - \ell f^*$ and $v_{f,f^*} = (r_{f,f^*})^2$

• Generalizes Tsybakov condition: $f^*$ does not need to be Bayes act, loss does not need to be 0/1
Bernstein Condition

• Fix a DP \((P, \ell, \mathcal{F})\) with (for now) bounded loss
• DP satisfies the \((C, \alpha)\)-Bernstein condition if there exists \(C > 0, \alpha \in [0, 1]\), such that for all \(f \in \mathcal{F}\)

\[
E[v_{f, f^*}] \leq C \cdot (E[r_{f, f^*}])^\alpha
\]

where we set \(r_{f, f^*} = \ell_f - \ell_{f^*}\) and \(v_{f, f^*} = (r_{f, f^*})^2\)

• Suppose data are i.i.d. and the \((C, \alpha)\)-Bernstein condition holds. Then...
Under Bernstein($\mathcal{C}, \alpha$)

- Empirical Risk minimization satisfies, with high prob*,

$$
\mathbb{E}[r_{\hat{f}_{\text{ERM}}, f^*}] = O \left( \left( \frac{\log |\mathcal{F}|}{T} \right)^{\frac{1}{2-\alpha}} \right)
$$

- $\alpha = 0$: condition trivially satisfied, get minimax rate

$$
O(1/\sqrt{T})
$$

- $\alpha = 1$: nice case (Massart condition), get ‘log-loss’ rate

$$
O(1/T)
$$
Under Bernstein\((\mathcal{C}, \alpha)\)

- \(\eta\) — “Bayes” MAP satisfies, with high prob*,

\[
\mathbb{E}[r_{\hat{f}_{\text{MAP}}, f^*}] = O \left( \left( \frac{-\log \pi(f^*)}{T} \right)^{\frac{1}{2-\alpha}} \right)
\]

- This requires setting “learning rate” \(\eta\) in terms of \(\alpha\) and \(T\)!
- \(\alpha = 0\): slow rate \(O(1/\sqrt{T})\); \(\alpha = 1\): fast rate \(O(1/T)\)
GOAL: Sequential Bernstein

- $\eta$ — “Bayes” MAP satisfies, with high prob*,

$$\mathbb{E}[r_{\hat{f}_{\text{MAP}}, f^*}] = O\left(\left(-\frac{\log \pi(f^*)}{T}\right)^{\frac{1}{2-\alpha}}\right)$$

- GOAL: design ‘sequential Bernstein condition’ and accompanying sequential prediction algorithm s.t.

1. cumulative regret always satisfies, for all $f^*$, all sequences

$$T^{-1} \cdot R_{\text{ALG}, f^*} = O\left(\left(-\frac{\log \pi(f^*)}{T}\right)^{\frac{1}{2-\alpha}}\right)$$

2. if condition holds, it also satisfies, with high prob*

$$T^{-1} \cdot R_{\text{ALG}, f^*} = O\left(\left(-\frac{\log \pi(f^*)}{T}\right)^{\frac{1}{2-\alpha}}\right)$$
GOAL: Sequential Bernstein

• **GOAL:** design ‘sequential Bernstein condition’ and accompanying sequential prediction algorithm s.t.
  1. cumulative regret always satisfies, for all $f^*$, all sequences

$$ R_{ALG,f^*} = O \left( T^{\frac{1}{2}} \cdot \left( - \log \pi(f^*) \right)^{\frac{1}{2}} \right) $$

  2. if condition holds, it also satisfies, with high prob*

$$ R_{ALG,f^*} = O \left( T^{\frac{1-\alpha}{2-\alpha}} \cdot \left( - \log \pi(f^*) \right)^{\frac{1}{2-\alpha}} \right) $$
DREAM

- **DREAM**: design ‘sequential Bernstein condition’ and accompanying sequential prediction algorithm s.t.
  1. cumulative regret always satisfies, for all \( f^* \), all sequences
    \[
    R_{\text{ALG}, f^*} = O \left( T^{\frac{1}{2}} \cdot (- \log \pi(f^*))^{\frac{1}{2}} \right)
    \]
  2. if condition holds for given sequence, then cumulative regret satisfies, for that sequence:
    \[
    R_{\text{ALG}, f^*} = O \left( T^{\frac{1-\alpha}{2-\alpha}} \cdot (- \log \pi(f^*))^{\frac{1}{2-\alpha}} \right)
    \]
GOAL: Sequential Bernstein

- **GOAL:** design ‘sequential Bernstein condition’ s.t.
  1. for all $f^*$, all sequences

$$R_{\text{ALG}, f^*} = O \left( T^{\frac{1}{2}} \cdot (- \log \pi(f^*))^{\frac{1}{2}} \right)$$

  2. if condition holds, it also satisfies, with high $\text{prob}^*$,

$$R_{\text{ALG}, f^*} = O \left( T^{\frac{1-\alpha}{2-\alpha}} \cdot (- \log \pi(f^*))^{\frac{1}{2-\alpha}} \right)$$

Approach 1: define seq. Bernstein as standard Bernstein+i.i.d.
Even then none of the standard algorithms achieve this...
*With one (?) exception!*
Today: Three Things To Tell You

1. Nifty Reformulation of Fast Rate Conditions in Statistical Learning

2. Do this via new concept: ESI

3. Precise Analogue of Bernstein Condition for Fast Rates in Individual Sequence Setting
   – ...and algorithm that achieves these rates!
Exponential Stochastic Inequality (ESI)

• For any given $\eta > 0$ we write $X \leq_{\eta}^* \epsilon$ as shorthand for

\[ \mathbb{E}[e^{\eta X}] \leq e^{\eta \epsilon} \]

• $X \leq_{\eta}^* \epsilon$ implies, via Jensen,

\[ \mathbb{E}[X] \leq \epsilon \]

• $X \leq_{\eta}^* \epsilon$ implies, via Markov, for all $A$,

\[ P(X \geq \epsilon + A) \leq e^{-\eta A} \]
ESI-Example

• Hoeffding’s Inequality: suppose that $X$ has support $[-1,1]$, and mean 0. Then

$$X \leq_{\eta} \mathbb{E}[X] + \frac{\eta}{2}$$
ESI – More Properties

- For i.i.d. rvs $X, X_1, ..., X_T$ we have

$$X \leq^*_\eta \epsilon \Rightarrow \sum_{t=1}^T X_t \leq^*_\eta T \cdot \epsilon$$

- For arbitrary rvs $X, Y$ we have

$$X \leq^*_\eta a ; Y \leq^*_\eta b \Rightarrow X + Y \leq^*_\eta/2 a + b$$
Bernstein in ESI Terms

Most general form of Bernstein condition: for some nondecreasing function $s : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$:

$$\forall f \in \mathcal{F} : \mathbb{E}[v_{f,f^*}] \leq s(\mathbb{E}[r_{f,f^*}])$$
Bernstein in ESI Terms

• Most general form of Bernstein condition: for some nondecreasing function $s : \mathbb{R}_0^+ \to \mathbb{R}_0^+$:

$$\forall f \in \mathcal{F} : \mathbb{E}[v_{f,f^{*}}] \leq s(\mathbb{E}[r_{f,f^{*}}])$$

• Van Erven et al. (2015) show this is equivalent to having

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \ell_{f^{*}} - \ell_f \leq u(\epsilon) \epsilon$$

for some nondecreasing function $u : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with

$$u(x) \asymp \frac{x}{s(x)}$$
U-Central Condition

Van Erven et al. (2015) show Bernstein condition is equivalent to the existence of increasing function $u : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ such that for some $f^* \in \mathcal{F}$:

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad \ell f^* - \ell f \leq^*_{u(\epsilon)} \epsilon$$

They term this the $u$-central condition.
Van Erven et al. (2015) show Bernstein condition is equivalent to the existence of increasing function $u : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ such that for some $f^* \in \mathcal{F}$:

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad \ell f^* - \ell f \leq^*_u(\epsilon) \epsilon$$

They term this the \textit{u-central condition} – can also be related to \textit{mixability}, \textit{exp-concavity}, \textit{JRT-condition}, condition for well-behavedness of \textit{Bayesian} inference under misspecification.
U-Central Condition

- Van Erven et al. (2015) show Bernstein condition is is equivalent to the existence of increasing function $u : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ such that for some $f^* \in \mathcal{F}$:

  $$\forall f \in \mathcal{F}, \epsilon \geq 0 : \ell f^* - \ell f \leq^*_{u(\epsilon)} \epsilon$$

They term this the $u$-central condition
- can also be related to mixability, exp-concavity, JRT-condition, condition for well-behavedness of Bayesian inference under misspecification
- for unbounded losses, it becomes different (and better!) than Bernstein condition – it is one-sided
Three Equivalent Notions for Bounded Losses

• U-central condition in terms of regret:

\[ \forall f \in \mathcal{F}, \epsilon \geq 0 : \quad -r_{f,f^*} \leq^* u(\epsilon) \epsilon \]

.....or equivalently (extending notation):

\[ \forall f \in \mathcal{F}, \epsilon \geq 0 : \quad 0 \leq^* u(\epsilon) r_{f,f^*} + \epsilon \]
Three Equivalent Notions for Bounded Losses

- U-central condition in terms of regret: with $\eta := u(\epsilon)$

\[ \forall f \in \mathcal{F}, \epsilon \geq 0 : \quad 0 \leq^*_\eta r_{f,f^*} + \epsilon \]

- For bounded losses, this turns out to be equivalent to: for some appropriately chosen $C_1, C_2$ with $\eta_\epsilon := C_1 u(\epsilon)$:

\[ \forall f \in \mathcal{F}, \epsilon \geq 0 : \quad C_2 \cdot \eta_\epsilon \cdot v_{f,f^*} \leq^*_\eta r_{f,f^*} + \epsilon \]
Three Equivalent Notions for Bounded Losses

• U-central condition in terms of regret: with $\eta := u(\epsilon)$

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad 0 \leq^* r_{f,f^*} + \epsilon$$

• For bounded losses, this turns out to be equivalent to: for some appropriately chosen $C_1, C_2$ with $\eta_\epsilon := C_1 u(\epsilon)$:

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad C_2 \cdot \eta_\epsilon \cdot \nu_{f,f^*} \leq^* r_{f,f^*} + \epsilon$$

• More similar to original Bernstein condition. However, condition is now in ‘exponential’ rather than ‘expectation’ form
Today: Three Things To Tell You

1. **Nifty Reformulation** of Fast Rate Conditions in Statistical Learning

2. Do this via new concept: **ESI**

3. Precise Analogue of Bernstein Condition for Fast Rates in Individual Sequence Setting
   
   – ...and algorithm that achieves these rates!
Suppose that \( u \)-central condition holds (i.e. \( x / u(x) \) – Bernstein holds), and data are i.i.d.

Then by generic property of ESI, with \( \eta_\epsilon = C_1 \cdot u(\epsilon) \),

\[
\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad C_2 \cdot \eta_\epsilon \cdot V_{f,f^*} \leq_{\eta_\epsilon} R_{f,f^*} + T \cdot \epsilon
\]

where \( R_{f,f^*} = \sum_{t=1}^{T} (\ell_{f,t} - \ell_{f^*,t}) \)

\[
V_{f,f^*} = \sum_{t=1}^{T} (\ell_{f,t} - \ell_{f^*,t})^2
\]
T-fold U-Central Condition

- Under $u$-central cond. and iid data, with $\eta_\epsilon = C_1 \cdot u(\epsilon)$:

  $$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad C_2 \cdot \eta_\epsilon \cdot V_{f,f^*} \leq_{\eta_\epsilon} R_{f,f^*} + T \cdot \epsilon$$

  but also for every learning algorithm $\text{ALG} : \bigcup_{t \geq 0} \mathcal{L}_t \rightarrow \mathcal{F}$

  $$C_2 \cdot \eta_\epsilon \cdot V_{\text{ALG},f^*} \leq_{\eta_\epsilon} R_{\text{ALG},f^*} + T \cdot \epsilon$$

with

$$R_{\text{ALG},f^*} = \sum_{t=1}^{T} (\ell_{\text{ALG},t} - \ell_{f^*,t})$$

$$V_{\text{ALG},f^*} = \sum_{t=1}^{T} (\ell_{\text{ALG},t} - \ell_{f^*,t})^2$$
Cumulative U-Central Condition

- Under $u$-central cond. and iid data, with $\eta_\epsilon = C_1 \cdot u(\epsilon)$:

$$\forall f \in \mathcal{F}, \epsilon \geq 0 : \quad C_2 \cdot \eta_\epsilon \cdot V_{f,f^*} \leq_{\eta_\epsilon} R_{f,f^*} + T \cdot \epsilon$$

but also for every learning algorithm $\text{ALG} : \bigcup_{t \geq 0} \mathcal{L}_t \rightarrow \mathcal{F}$

$$C_2 \cdot \eta_\epsilon \cdot V_{\text{ALG},f^*} \leq_{\eta_\epsilon} R_{\text{ALG},f^*} + T \cdot \epsilon$$

This condition may of course also hold for non-i.i.d. data. It is the condition we need, so we term it the cumulative u-central condition.
Hedge with Oracle Learning Rate

• Hedge with learning rate $\eta$ achieves regret bound, for all $f^* \in \mathcal{F}$

$$R_{\text{HEDGE}}(\eta), f^* \leq C_0 \cdot \eta \cdot V_{\text{ALG}, f^*} + \frac{-\log \pi(f^*)}{\eta}$$

• We assume cumulative $u$-central condition for some $u$. For simplicity assume $u(x) \simeq x^\beta$; then:

$$\forall \epsilon \geq 0, \eta = C_1 \cdot \epsilon^\beta: \quad C_2 \cdot \eta \cdot V_{\text{ALG}, f^*} \leq \eta R_{\text{ALG}, f^*} + T \cdot \epsilon$$

and even for some other constant

$$\forall \epsilon \geq 0, \eta = C'_1 \cdot \epsilon^\beta: \quad C_0 \cdot \eta \cdot V_{\text{ALG}, f^*} \leq \frac{1}{2} R_{\text{ALG}, f^*} + \frac{T}{2} \cdot \epsilon$$
Hedge with Oracle Learning Rate

• Combining we get $\forall \epsilon \geq 0, \eta = C_1' \cdot \epsilon^\beta$

$$\frac{1}{2} R_{\text{HEDGE}}(\eta), f^* \leq^* T \cdot \epsilon / 2 + \frac{-\log \pi(f^*)}{\eta}$$

• We can set $\epsilon$ (or eqv. $\eta$) as we like. Best possible bound achieved if we make sure all terms are of same order, i.e. we set at time $T$,

$$T \cdot \epsilon / 2 = \frac{-\log \pi(f^*)}{\eta}$$

• and then $\eta_T \asymp \left( \frac{-\log \pi(f^*)}{T} \right)^{\frac{\beta}{1+\beta}}$ and

$$R_{\text{HEDGE}}(\eta_T), f^* \leq^* \eta_T / 2 \cdot C \cdot T^{\frac{\beta}{1+\beta}} \cdot ( - \log \pi(f^*) )^{\frac{1}{1+\beta}}$$
Squint without Oracle Learning Rate!

- Hedge achieves ESI- (!)-bound
  \[ R_{\text{HEDGE}}(\eta), f^* \leq_{\eta/2}^* C \cdot T^{\frac{\beta}{1+\beta}} \cdot (\log \pi(f^*))^{\frac{1}{1+\beta}} \]

  ...but needs to know \( f^*, \beta \) and \( T \) to set learning rate!

- **Squint** (Koolen and Van Erven ’15)
  - achieves same bound without knowing these!
  - Gets bound with \( \beta = 0 \) automatically for individual sequences

- What about **Adanormalhedge**? (Luo & Shapire ‘15)
Dessert: Easy Data Rather than Distributions

• We are working with algorithms such as Hedge and Squint, designed for individual, nonstochastic sequences
• Yet condition is stochastic
• Does there exist nonstochastic analogue?
• Answer is yes:
Non-Stochastic Inequality

Suppose $u$-cumulative central condition holds for some $u$. Using Martingale theory one shows that this also implies the following:

• fix a countable, otherwise arbitrary set $\mathcal{A}$ of learning algorithms.

• Fix a decreasing sequence $\epsilon_1, \epsilon_2, \ldots$ and set corresponding $\eta_1 = u(\epsilon_1), \eta_2 = u(\epsilon_2), \ldots$

• Then we have with probability 1: for every $\text{ALG} \in \mathcal{A}$ there exists $C$ such that

$$\forall T > 0 : C_2 \cdot \eta_T \cdot V_{\text{ALG}, f^*} \leq R_{\text{ALG}, f^*} + T \cdot (\log \log T) \cdot \epsilon_T + C$$
Individual Sequence Condition

Hence we define:

(we only give special case with $u(x) = x^\beta$ here)

An **individual sequence** satisfies the $u$-fast rate condition relative to countable set of learning algorithms $\mathcal{A}$ and constants $\{C_{\text{ALG}} : \text{ALG} \in \mathcal{A}\}$ if there exists $f^*$ such that for all $T > 0$, for all $\text{ALG} \in \mathcal{A}$, with

$$\eta_T = \left( \frac{-\log \pi(f^*)}{T} \right)^{1/1+\beta}$$

$$\epsilon_T = \left( \frac{-\log \pi(f^*)}{T} \right)^{1/1+\beta}$$

we have

$$C_2 \cdot \eta_T \cdot V_{\text{ALG},f^*} \leq R_{\text{ALG},f^*} + T \cdot (\log \log T) \cdot \epsilon_T + C_{\text{ALG}}$$
Conclusion

- If a \textit{sequence} satisfies \( u \)-fast rate condition, then Hedge (with oracle) and Squint (without oracle) both achieve desired regret bound.
- We’ve removed all stochastics!
  - Similar idea used by György and Szepesvári in this workshop!
- Notion implies a (very close!) analogy to \textit{Martin-Löf randomness}

Van Erven, G. Mehta, Reid, Williamson

\textit{Fast Rates in Statistical and Online Learning.}
JMLR Special Issue in Memory of A. Chervonenkis, Oct. 2015
Iets zeggen over: L* bound, unbounded losses, mixability, JRT, exp-concavity, ....
Tell Csaba, Peter B, Philippe
\( \eta \leq u(\epsilon) \), maar ook met \( \eta = u(\epsilon) \)
Star means...